
BrainLang: Predicting Heard Language through fMRI Text Decoding

Riad Dajani

Johns Hopkins University
AS.050.637: RTM by Dr. Li
rdajani2@jh.edu

1 Introduction

In the interdisciplinary field of neuroscience and computational technology, Brain-Computer Interfaces (BCIs) represent a significant advance for exploring and enhancing human cognitive capabilities. This research project delves into a pivotal aspect of BCIs—decoding fMRI data into coherent language outputs [6]. This capability not only promises to improve communication aids for individuals with speech impairments [1] but also extends the potential of BCIs to enhance cognitive bandwidth, allowing for new ways to interact with and control our environment through decoded neural activity [1, 5, 6]. By employing sophisticated neural network models, this research seeks to determine how effectively language processing can be mapped from brain signals, moving us closer to harnessing the full potential of our cognitive processes in real-time applications.

2 Background

Decoding neural activity into actionable language constructs has been bolstered by key advancements in functional Magnetic Resonance Imaging (fMRI) and deep learning. Notably, it has been demonstrated how GPT-2 could be utilized to derive meaningful neural encodings from fMRI data during narrative listening [5], highlighting the potential of deep learning to interpret complex brain signals. Moreover, the use of LSTM models to construct interpretable, multi-timescale representations, which excel in capturing the temporal dynamics has been proven extremely effective for decoding fMRI responses [4]. Furthermore, the efficacy of meaningfully reconstructing continuous language from brain activity in a non-invasive manner has been bolstered [1], setting a precedent for both detailed and practical decoding applications. Together, these studies underscore the feasibility and the complexity of extracting coherent language from neural signals and frame the scientific challenge that this research aims to address—enhancing the precision and utility of neural decoding for both therapeutic and augmentative applications.

Following, this project aims to develop a sophisticated model that utilizes BiLSTM networks [8] and UMAP dimensionality reduction [7] to predict and decode sequences of text from fMRI data during a

natural language listening task [3]. By integrating these methodologies, the study seeks to overcome the existing limitations in neural decoding accuracy and temporal resolution [1]. The methods section will detail the dataset involving fMRI data from subjects listening to narrative stories, the preprocessing steps, and the modeling techniques employed. Following this, the results section will present the outcomes of the model's performance, including comparisons with established benchmarks. Finally, the discussion will interpret these results in the context of their implications for enhancing cognitive bandwidth and communicational capabilities through brain-computer interfaces, highlighting potential applications and suggesting avenues for future research.

3 Methods

3.1 Dataset Description

The project utilizes the dataset from "A natural language fMRI dataset for voxelwise encoding models" as detailed by Amanda LeBel et al. (2023). This comprises functional Magnetic Resonance Imaging (fMRI) responses from eight participants (three females, five males, ages 21-34) who listened to 27 complete natural narrative stories sourced from The Moth, each lasting 10 to 15 minutes [3]. These narratives were delivered without scripts, providing a natural listening experience that closely mimics everyday language comprehension.

The fMRI data provided is already preprocessed and was used in this project. Pre-processing steps included motion correction using the FMRIB Linear Image Registration Tool (FLIRT), cross-run alignment, and the removal of low-frequency voxel response drifts using a 2nd order Savitzky-Golay filter [3]. Data was collected using a gradient-echo EPI sequence with a repetition time (TR) of 2.0 seconds [3].

In addition to the fMRI data, the dataset includes manually transcribed text of the narratives, aligned with the audio recordings using the Penn Phonetics Lab Forced Aligner [9]. This alignment allows for precise temporal matching of text data to fMRI responses, essential for developing accurate voxelwise encoding models.

3.2 Experimental Paradigm

The experimental setup involved playing the narrative stories through headphones in an fMRI scanner, ensuring clear audio reception while minimizing external noise. Participants were instructed to remain still and attentive to minimize motion artifacts [3]. This setup facilitated the synchronization of spoken words with fMRI data, enabling the precise analysis of neural correlates of language processing during the listening task.

3.3 Data Preprocessing

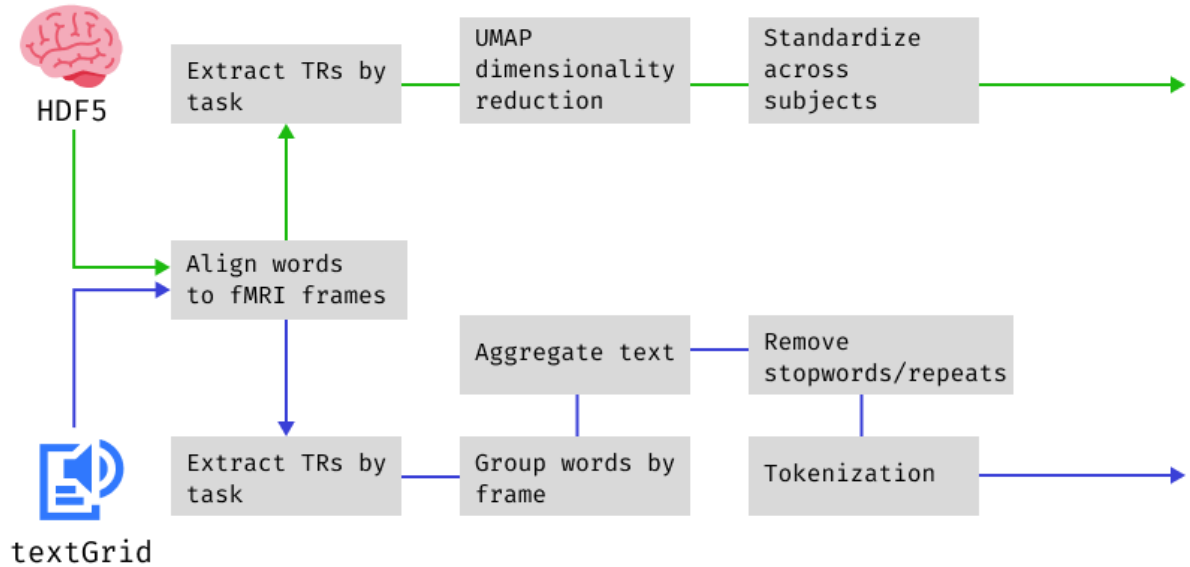
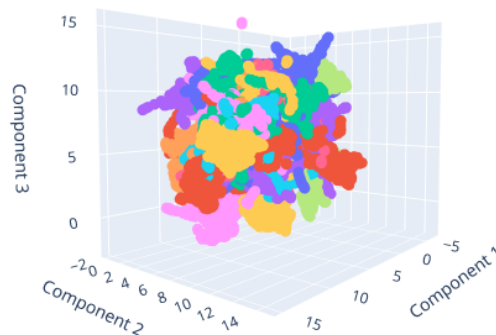


Figure 1: Data Preparation and Preprocessing Prior to Model Use - Data Preprocessing

fMRI Data Preprocessing

This study utilizes preprocessed fMRI data provided in the dataset described by Amanda LeBel et al. (2023). These preprocessed scans included motion-corrected and cross-run aligned MRI data, prepared using established protocols to ensure high data quality and consistency [3]. Beyond the provided preprocessing, to further reduce the dimensionality of the fMRI data and enhance model performance, Uniform Manifold Approximation and Projection (UMAP) [7] was applied. UMAP reduced the data dimensions while preserving the essential structural and functional relationships, making the high-dimensional data more tractable for neural network analysis (see figure 2). The parameters set for UMAP included a neighbor count of 50, a minimum distance of 0.1, and an output dimensionality of 5.

UMAP Embeddings of fMRI Data in 3D



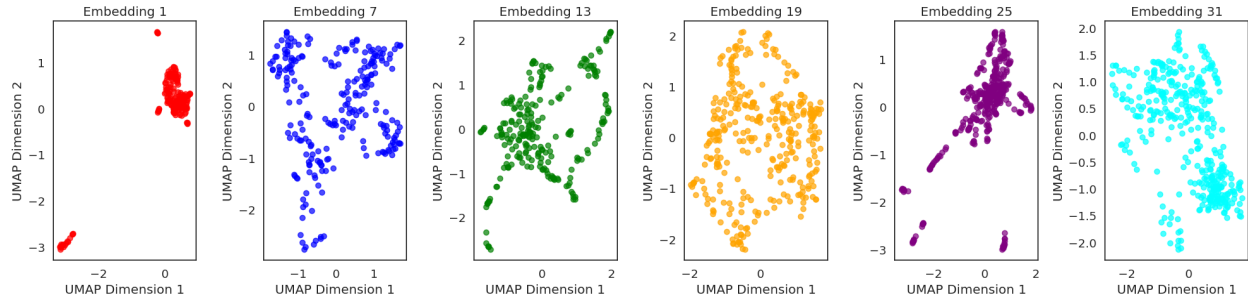


Figure 2: 3D (top) and 2D samples (bottom) of UMAP Embeddings of fMDI Data

Text Data Preprocessing:

The corresponding text data is processed by aggregating the transcriptions aligned with each fMRI frame (Figure 3), removing stop words and repeated words, and tokenizing the resulting sequences using the Keras Tokenizer. The tokenized sequences are padded to a uniform length. The decoder inputs are prepared by shifting the target sequences by one time step, allowing the model to learn to predict the next word given the previous words. The decoder targets are one-hot encoded to match the model's output dimensionality.

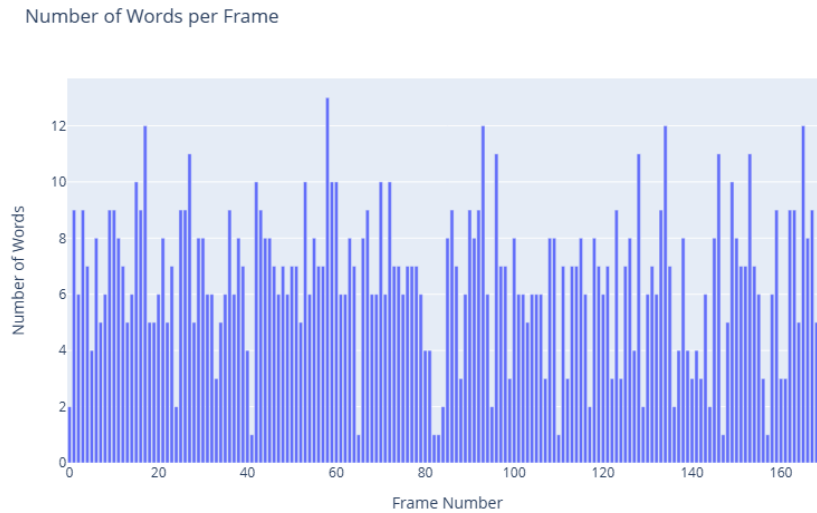


Figure 3: Distribution of Words Per Frame for Sample Task - Data Preprocessing

Standardization and Aggregation:

Following the dimensionality reduction with UMAP, the embeddings were standardized using a StandardScaler to ensure that the neural network model received data with zero mean and unit variance. This standardization is crucial for the effective training of deep learning models. The text data was similarly prepared, with sequences aggregated and tokenized to create a uniform input for the decoding stage of the model.

3.3.1 Supplemental Description of UMAP

Uniform Manifold Approximation and Projection (UMAP) is a dimensionality reduction technique that preserves the local and global structure of data [7]. Developed by McInnes, Healy, and Melville in 2018, UMAP is particularly effective for complex datasets like fMRI, where maintaining the integrity of neural patterns is crucial. It is used in this study to facilitate the neural network's ability to decode fMRI data into text, enhancing model performance by simplifying the input data while retaining essential information. It also helps in addressing the low signal-to-noise ratio (SNR) issue highlighted by Tang et al. (2023), as it enhances the specificity of model features potentially increasing decoding performance [1, 7].

4 Modeling/Experiments and Analysis

4.1 Model Architecture

The BrainLang model employs an architecture specifically designed to decode spoken language from the reduced fMRI data (see figure 1 for more details on data processing). At its core, the model utilizes a sequence-to-sequence (seq2seq) framework, which consists of an encoder and a decoder (figure 4).

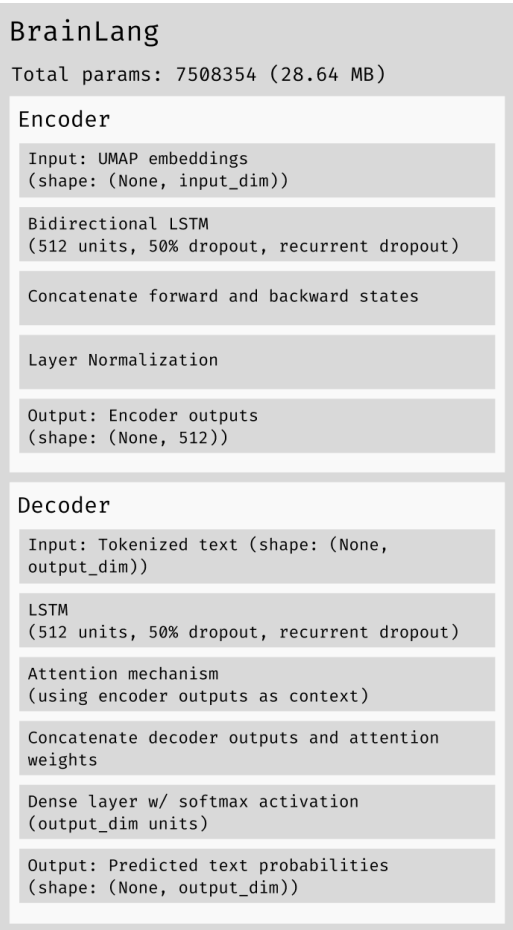


Figure 4: Visualization of BranLang Model - Model Architecture

4.2 Encoder/Decoder Stacks

Encoder: The encoder component of the BrainLang model is designed to process input sequences derived from UMAP-reduced fMRI data. It utilizes a Bidirectional LSTM (BiLSTM) layer, which enhances the model's ability to capture contextual relationships in the data by processing the sequence both forwards and backwards [8]. This dual-direction processing allows the model to gather comprehensive temporal insights, crucial for understanding complex patterns in fMRI signals related to language comprehension [4].

Jain et al. (2020) suggested that traditional LSTM models might not effectively separate the temporal dynamics across multiple timescales, leading to a potential loss in capturing nuanced temporal patterns [4]. As such, the bidirectional approach was employed as it ensures a more comprehensive understanding of the temporal context, crucial for decoding the complex narrative-driven fMRI data.

The outputs from both directions are then concatenated to form a unified representation, which is further refined by dense layers to prepare a consolidated state for the decoder. This ensures that all relevant temporal features are captured and effectively integrated before being passed on to the decoder.

Decoder: The decoder takes the processed states from the encoder and aims to generate the corresponding textual output. It begins with an LSTM layer that receives initial states from the encoder, ensuring continuity in capturing the temporal dynamics of the input sequence. This LSTM processes the decoder inputs, which are shifted versions of the target sequences, facilitating the prediction of the next word in the sequence based on previous words (teacher forcing).

The outputs of the LSTM are then passed to an attention mechanism which focuses processing on the most relevant parts of the neural sequence, mimicking the cognitive prioritization of brain regions when processing specific words or concepts. The model used by Tang et al. (2023), provides grounds for the need to selectively focus on relevant features from the encoder outputs, which is critical for decoding tasks in brain-computer interfaces [1]. The attention layer in the BrainLang model can dynamically weigh the importance of different inputs across the sequence, enhancing the model's ability to decode neural signals into meaningful predictions

The final output is generated through a dense layer that translates the processed signals into probabilities across a large vocabulary of 4,407 possible words. In all, this setup allows the decoder to predict the most likely next word in the sequence, effectively translating the brain's encoded language signals into readable text.

4.3 Addressing Challenges

When designing BrainLang, the teachings and key challenges from previous papers related to decoding fMRI data for language processing were prioritized. This included addressing low signal-to-noise ratios, and non-cooperative contexts [1, 4, 5].

Tang et al. (2023) identified that low SNR and model mis-specification can significantly impair the quality of decoded information from fMRI data [1]. With this, the BiLSTM layer enhances the handling of noisy data by focusing on relevant temporal features from both past and future contexts. The attention mechanism further refines this process by selectively enhancing the model's focus on the most relevant features [2], ensuring that decoding errors caused by misaligned or suboptimal feature representations are minimized.

Tang et al. also noted a significant drop in decoding performance when subjects resisted by engaging in different cognitive tasks [1]. While the model does not directly address the involuntary nature of subject cooperation, the inclusion of an attention mechanism enhances its ability to detect subtler signals that might be present in non-cooperative contexts [2]. This feature could potentially allow the model to capture useful neural information even when subjects are not actively cooperating or when their cognitive focus shifts, offering a valuable tool for scenarios where subject compliance cannot be guaranteed.

4.4 Training

The training involved preparing encoder input data from UMAP-reduced fMRI sequences and decoder input data through a technique known as 'teacher forcing', where sequences are shifted by one time step. The model was trained on these inputs using a batch size of 32 for up to 100 epochs, with early stopping and learning rate reduction implemented to prevent overfitting and ensure convergence. Specifically, training was halted if no improvement in validation loss was observed for 10 epochs, and the learning rate was reduced by a factor of 0.5 under similar conditions, with a lower bound set at 0.0001.

5 Result Figures

5.1 Model Accuracy/Loss

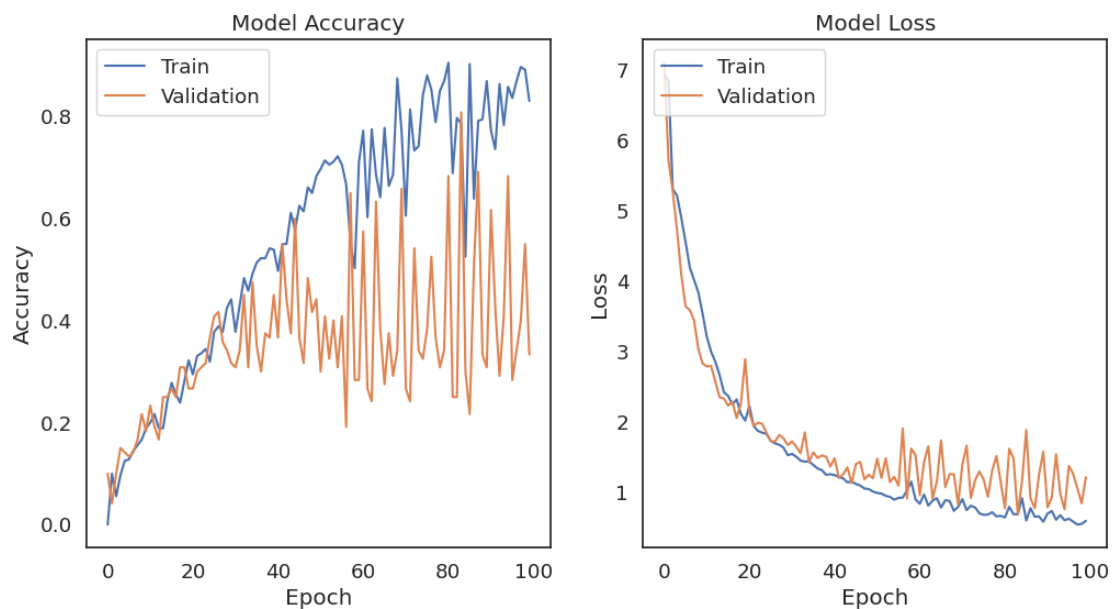


Figure 5: Model Accuracy and Loss from a Training Run (4 tasks, 8 subjects)

5.2 Prediction Accuracy

Predicted:

loud like he always has it and and know know that now
there there there my my snuggled snuggled snuggled up
by his his his his his his his grandfather

Actual:

loud like he always has it and i know that now and
there was my son just snuggled up by his side and
trading wild animal facts with his grandfather

Figure 6: Output Sample 8 from Model After Training (green = match, red = false, Accuracy: 0.3167)

6 Discussions

6.1 Summary and Interpretation of Results

The BrainLang model has demonstrated a foundational capacity to decode sequences of heard language from fMRI data. Analyzing specific predictions, the model exhibited a preliminary understanding of the language structure, though it showed noticeable repetition and substitution errors. For instance, in Sample 8, the phrase "loud like he always has" was accurately captured, but it faltered with repetitive and misplaced words like "know now and there there my son snuggled snuggled." (see figure 6). Similarly, Sample 9 maintained parts of the sentence structure but struggled with repetition and coherence towards the end.

The achieved accuracy of 31.67% (figure 5) highlights the model's initial success in recognizing and generating coherent language elements from complex brain signals, though it also underscores the challenges inherent in this task. This level of accuracy, while modest, is a significant step in decoding continuous spoken language from fMRI data—a complex endeavor given the high dimensionality and noise levels associated with fMRI datasets [1]. The results reflect the model's ability to grasp basic linguistic patterns and contextual cues but also reveal the limitations of current technologies in fully capturing the nuanced dynamics of human language processing.

These outcomes are particularly meaningful when considering the novel application of bidirectional LSTM networks and attention mechanisms in interpreting temporally dynamic and spatially distributed neural signals [4]. The comparison of predicted versus actual texts illustrates the potential and current constraints of our approach, emphasizing areas for enhancement in both the model architecture and training processes to improve future decoding accuracy and reliability [1].

6.2 Strengths, Limitations, and Comparisons

A key strength of the model is its integration of bidirectional LSTM layers, which allow it to account for both past and future contextual information [8] within the fMRI data, in contrast to the transformer approach in Tang et al. (2023) which focused solely on discrete phrase decoding [1]. Moreover, BrainLang attempts a more challenging task of continuous language decoding, which may explain the lower accuracy but also suggests greater potential utility. Unexpectedly, the model's performance did not plateau early as typical in high-dimensional settings (see figure 5), suggesting that further epochs could yield improved results. This methodological deviation may support more robust semantic decoding from noisy fMRI data.

However, limitations persist in the form of high computational demands and the need for extensive data preprocessing to achieve optimal decoding accuracy. An unexpected finding was the model's ability to maintain reasonable accuracy even with reduced data (fewer tasks), suggesting potential applications in more portable, less resource-intensive brain imaging technologies, akin to findings from Tang et al. about using lower-resolution fMRI data effectively [1].

6.3 Looking Ahead

Future research could explore the integration of multimodal data sources, such as substituting fMRI with iEEG, to enhance decoding accuracy and speed [1], as suggested by the successful results with low-temporal fMRI in Tang et al. (2023). Additionally, refining the attention mechanisms within our LSTM framework could further address the challenges of decoding [2] from non-cooperative subjects, a significant concern noted in both studies. Furthermore, similar datasets using stereotactic electroencephalography (sEEG) exist, and could help circumvent both temporal and spatial resolution issues. This approach seems more grounded, especially considering recent trends in consumer-friendly BCIs like Neuralink [10]. Moreover, recent advancements in transformers offer a promising avenue for developing lightweight, robust, and quick methods for decoding with higher accuracy [1, 5, 6].

References

- [1] J. Tang, A. LeBel, S. Jain, and A. G. Huth, “Semantic reconstruction of continuous language from non-invasive brain recordings,” *Nature*, Sep. 2022, doi: <https://doi.org/10.1101/2022.09.29.509744>.
- [2] A. Vaswani *et al.*, “Attention Is All You Need,” 2017. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
- [3] A. LeBel *et al.*, “A natural language fMRI dataset for voxelwise encoding models,” *Scientific Data*, vol. 10, no. 1, p. 555, Aug. 2023, doi: <https://doi.org/10.1038/s41597-023-02437-z>.
- [4] S. Jain, V. A. Vo, Shivangi Mahto, A. LeBel, J. S. Turek, and A. Huth, “Interpretable multi-timescale models for predicting fMRI responses to continuous natural speech,” *NeurIPS*, vol. 33, pp. 13738–13749, Jan. 2020.
- [5] A. G. Russo, A. Ciarlo, S. Ponticorvo, F. Di Salle, G. Tedeschi, and F. Esposito, “Explaining neural activity in human listeners with deep learning via natural language processing of narrative text,” *Scientific Reports*, vol. 12, no. 1, p. 17838, Oct. 2022, doi: <https://doi.org/10.1038/s41598-022-21782-4>.
- [6] Y. Yang, Y. Duan, Q. Zhang, R. Xu, and H. Xiong, “Decode Neural signal as Speech,” *arXiv (Cornell University)*, Mar. 2024, doi: <https://doi.org/10.48550/arxiv.2403.01748>.
- [7] L. McInnes, J. Healy, and J. Melville, “UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction,” *arXiv.org*, Sep. 17, 2020. [https://arxiv.org/abs/1802.03426#:~:text=UMAP%20\(Uniform%20Manifold%20Approximation%20and](https://arxiv.org/abs/1802.03426#:~:text=UMAP%20(Uniform%20Manifold%20Approximation%20and)
- [8] Z. Huang, W. Xu, and K. Yu, “Bidirectional LSTM-CRF Models for Sequence Tagging,” *arXiv.org*, Aug. 09, 2015. <https://arxiv.org/abs/1508.01991v1> (accessed Oct. 03, 2023).
- [9] P. Boersma and D. Weenink, “Praat: Doing Phonetics by Computer,” *Ear and Hearing*, vol. 32, no. 2, p. 266, Mar. 2011, doi: <https://doi.org/10.1097/aud.0b013e31821473f7>.
- [10] E. Musk, “An Integrated Brain-Machine Interface Platform With Thousands of Channels (Preprint),” *Journal of Medical Internet Research*, vol. 21, no. 10, Sep. 2019, doi: <https://doi.org/10.2196/16194>.

Supplementals

https://drive.google.com/drive/folders/1YJ4JE6iPYAuzqDhzakoFyiW2VoYIaPv_?usp=sharing